## Speedup of GEMV on A100 (Weight Quantize) cuBLAS-W<sub>FP16</sub>A<sub>FP16</sub> TensorRTLLM-WINTAAFP16 Marlin-W<sub>INT</sub>4A<sub>FP</sub>16 BitBLAS-W<sub>FP4</sub>A<sub>FP16</sub> FasterTransformer-W<sub>INT4</sub>A<sub>FP16</sub> vLLM-W<sub>INT4</sub>A<sub>FP16</sub> BitBLAS-W<sub>INT4</sub>A<sub>FP16</sub> BitBLAS-W<sub>INT2</sub>A<sub>INT8</sub> VS dnpaeds V0 V4 V5 V6 Shapes from LLM